ETS®

Listening. Learning. Leading.®

Guidelines for the Assessment of English Language Learners

Guidelines for the Assessment

of English Language Learners

Preface

The proper assessment of our nation's more than 5 million English Language Learners (ELLs) merits attention at all levels in our education systems. It is critically important that the array of content assessments taken by ELLs be fair and valid. That is no easy task, but it is key to improving educational opportunities for language-minority students.

Fortunately, Educational Testing Service has published this new comprehensive guide. It will be of great value to test developers, test administrators, educators, education policymakers and others. The 27-page *Guidelines for the Assessment of English Language Learners* is the latest in a series of research-based ETS publications that address quality issues as they relate to fairness and equity in testing.

ELLs are students who are still developing proficiency in English. They represent one in nine students in U.S. classrooms from pre-kindergarten through grade 12, but most are concentrated in the lower grades. Collectively, they speak about 400 languages, although approximately 80 percent are native speakers of Spanish. Persons of Asian descent — primarily speakers of Mandarin, Cantonese, Hmong and Korean — account for about 5 percent of the balance of the ELL population. While most of these students are found in large urban centers, many others live in concentrations in smaller communities.

English-language learners are concentrated in six states — Arizona, California, Texas, New York, Florida and Illinois. The ELL students in those six states account for more than 60 percent of the ELL population.

As principal author and Senior Research Scientist and Research Director John Young notes, "The U.S. federal government's No Child Left Behind legislation of 2001 has made the need to produce valid and fair assessments for ELLs a matter of pressing national concern. So we produced a framework to assist practitioners, educators, test developers and educators in making appropriate decisions on assessment of ELLs in academic content areas."

The No Child Left Behind Act, or NCLB, includes ELLs as one of the mandated subgroups whose test scores are used to determine whether schools and school districts throughout the United States are meeting goals for what the law refers to as "adequate yearly progress" (AYP) based on state-level performance standards established for their students.

Because almost all assessments measure language proficiency to some degree, the guidelines point out, ELLs may receive lower scores on content area assessments administered in English than they would if they took the same tests in a language in which they were proficient.

And that is why the new guide is so important: it helps educators assess students' mastery of subject matter while minimizing the role of the student's English proficiency in its measurement.

These guidelines are the latest in a series of actions that ETS has taken in recent years to support the pursuit of quality, fairness and accuracy in English-language learner assessments. One such program

was a 2008 symposium, "The Language Acquisition and Educational Achievement of English Language Learners," co-convened by ETS and the National Council of La Raza (NCLR).

NCLR Vice President for Education Delia Pompa shares my view that "ETS renders a great service in issuing these guidelines. They are a welcome and much needed addition to our collective knowledge following our ETS-NCLR ELL symposium last year, and will advance teaching and testing for ELL practitioners everywhere."

In commending ETS for this extremely valuable publication, I urge all ELL stakeholders to read it and take full advantage of its recommendations. All of our learners deserve the best opportunities we can provide. Fair and valid assessments are a key ingredient in that process.

Kenji Hakuta, Ph.D.
Lee L. Jacks Professor of Education
Stanford University

The *Guidelines for the Assessment of English-Language Learners* were authored by Mary J. Pitoniak, John W. Young, Maria Martiniello, Teresa C. King, Alyssa Buteux, and Mitchell Ginsburgh.

The authors would like to thank the following reviewers for their comments on an earlier version of this document: Jamal Abedi, Richard Duran, Kenji Hakuta, and Charlene Rivera. The authors would also like to acknowledge Jeff Johnson and Kim Fryer for the application of their excellent editing skills.

Contents

Introduction

Purpose and Audience

English language learners (ELLs)—students who are still developing proficiency in English—represent a large and rapidly growing subpopulation of students in U.S. classrooms. Accordingly, they are also a key group of students to consider when designing and administering educational assessments.

The guidelines in this document are designed to be of use to test developers, testing program administrators, psychometricians, and educational agencies as they work to ensure that assessments are fair and valid for ELLs. These guidelines focus on large-scale content area assessments[1] administered in the United States to students in grades K-12; however, many of the principles can be applied to other populations and other assessments.

These guidelines assume a basic knowledge of concepts related to educational testing. However, some sections may be more relevant to a given group of practitioners than others and some sections—for example, the section on statistical considerations—may call for familiarity with technical concepts.

We hope that these guidelines will encourage those involved with educational assessment to keep ELLs in mind throughout the development, administration, scoring, and interpretation of assessments, and that these guidelines will ultimately lead to better assessment practices for all students.

Readers should use these guidelines in conjunction with other ETS guidelines and resources that discuss best practices in testing. These ETS documents include, but are not limited to, the following:

- *ETS Standards for Quality and Fairness*
- *ETS Fairness Review Guidelines*
- *ETS International Principles for Fairness Review of Assessments*
- *ETS Guidelines for Constructed-Response and Other Performance Assessments*

Background

ELLs comprise a large and growing subpopulation of students. As of the 2006-07 school year, there were more than 5 million ELLs in prekindergarten (PK) to grade 12 classrooms, with a greater concentration of ELLs at the lower grade levels. These students represent 1 in 9 students in U.S. classrooms. They are projected to represent 1 in 4 students by the year 2025. In California, it is already the case that more than 25% of the students in grades PK-12 are ELLs. Nationally, about 80% of ELLs are native speakers of Spanish, but overall, ELLs speak about 400 different home languages.

[1] Within this document, the terms *assessment* and *test* are used interchangeably.

With the passage of the federal No Child Left Behind (NCLB) legislation in 2001—and with the increasing emphasis on accountability testing in general—the need to produce valid and fair assessments for ELLs has become a matter of pressing national concern. Under NCLB, the academic progress of ELLs is assessed in two ways:

(1) Under Title I, ELLs are one of the mandated subgroups whose test scores are used to determine whether schools and districts are meeting the goals for adequate yearly progress (AYP) based on state-level performance standards established for their students. ELLs are held to the same expectations as other subgroups regarding participation and attainment of proficiency on selected content area assessments (although ELL students are allowed a grace period during which the scores will not count).

(2) Under Title III, ELLs must also demonstrate progress in attaining English language proficiency.

The main purpose of these guidelines is to provide testing practitioners, as well as other educators, with a framework to assist in making appropriate decisions regarding the assessment of ELLs in academic content areas, including but not exclusively as specified under Title I. These guidelines do not focus on assessing English language proficiency, as defined under Title III.

Validity Issues in Assessing ELLs

As noted in the *ETS Standards for Quality and Fairness*, validity is one of the most important attributes of an assessment. Validity is commonly referred to as the extent to which a test measures what it claims to measure. For ELLs, as well as for all populations, it is critical to consider the degree to which interpretations of their test scores are valid reflections of the skill or proficiency that an assessment is intended measure.

Although there are several validity issues related to the assessment of ELLs, the main threat when assessing academic content areas stems from factors that are irrelevant to the construct—the skills or proficiency—being measured. The main goal of these guidelines is to minimize these factors—termed construct-irrelevant variance—and to ensure that, to the greatest degree possible, assessments administered to ELLs test only what they are intended to test.

Since almost all assessments measure language proficiency to some degree, ELLs may receive lower scores on content area assessments administered in English than they would if they took the same tests in a language in which they were proficient. For example, an ELL who has the mathematical skills needed to solve a word problem may fail to understand the task because of limited English proficiency. In this case, the assessment is testing not only mathematical ability, but

also English proficiency. If the construct of interest is *mathematical skill exclusive of language skills*, then it may be systematically inaccurate to base inferences about the academic content knowledge or skills of this student and other ELLs on the scores of tests delivered in English. This distinction can be complicated if the construct of interest is not merely *mathematical skill*, but rather *the ability to do mathematics within an English-medium classroom*. Please see the discussion of *Defining the Construct* later in the document.

To increase the validity of test score interpretations for ELLs in areas where English proficiency is not judged to be part of the construct of interest, testing practitioners can take a number of steps to maximize the degree to which the test scores reflect the individual's ability level in the content area being assessed, while minimizing the impact the student's level of English language proficiency has on those scores.

Caveats About Guidelines

Within these guidelines, we make many recommendations. In an ideal world, all of the recommendations could be implemented, but budgets and timeframes often require compromises. The realities of available funding and other resources will factor into decisions about which avenues to pursue; trade-offs between costs and benefits should be considered. Failure to follow all of this document's recommendations *will not* automatically make a test's scores *invalid*—but the possible impact on validity should always be considered.

Users of this document will need to make choices as to which recommendations to pursue, and they should consider factors such as the purpose of the test and the inferences to be made on the basis of the test scores. For example, if a test is used to make high-stakes decisions about a student, as would be the case for a high school graduation test, certain recommendations might carry more weight than if the test were used for remediation purposes. We encourage the reader to carefully consider each of the recommendations within the guidelines and to take into account the benefits of implementing them along with any challenges related to their execution.

In addition, as is noted in several sections of the guidelines, not all of the recommendations would work equally well with different types of ELLs, so users of this document must decide how to make the test accessible to most ELLs while minimizing difficulties that may be present for some ELL subgroups, such as those with very low levels of English language proficiency. Similarly, some test design features may benefit ELLs but prove challenging for other populations, such as students with visual impairments. In general, users of this document should carefully consider how to maximize accessibility for the greatest number of students both across and within subgroups.

Finally, although our recommendations are based on research and other documents relevant to the assessment of English language learners, we have chosen not to cite specific studies. Given the increased pace of work in this area, such references could be fast obsolete and are strictly speaking not required for understanding and implementing these guidelines. However, we have provided a bibliography at the end of the document that lists relevant articles for further exploration of the topic.

Overview of Guidelines

These guidelines are organized as follows: We start with definitions of the key terms used in this document. We then provide a general discussion of factors that can influence the assessment of ELLs. Next, we address developing assessment specifications, test items, and scoring criteria. This is followed by the sections on external test reviews and evaluating items. The last several sections of these guidelines focus on scoring constructed-response items, testing accommodations for ELLs, and using statistics to evaluate an assessment and scoring.

Key Terms

The following terms are used throughout the document:

- *Construct*—the skill or proficiency an assessment is intended to measure.
- *English language learner (ELL)*—in this document, a general term for students who are developing the English language proficiency needed to succeed in English-medium classrooms in U.S. schools.
- *Response*—any kind of performance to be evaluated as part of an assessment, including multiple-choice answers, short answers, extended answers, essays, presentations, demonstrations, or portfolios.
- *Rubric*—the scoring criteria, scoring guide, rating scale and descriptors, or other framework used to evaluate responses.
- *Task*—a specific test item, topic, problem, question, prompt, or assignment.
- *Testing accommodation*—any change to standardized testing conditions intended to make the test more fair and accessible for an individual or subgroup that does *not* change the construct being measured. These changes may include, but are not limited, to changes in the presentation of the assessment, the environment in which the assessment is administered, time allowed for the assessment, or additional materials or equipment to be used by students during the assessment.

- *Testing modification*—any change to standardized testing conditions that *does* change the construct being measured. For example, allowing a reading test to be read aloud to a student would be a modification if the construct being measured is decoding of text.
- *Testing variation*—an umbrella term referring to a change to standardized testing conditions; it may include either a testing accommodation or a testing modification.

As noted above, this document will use the term ELL to refer to students who are in the process of developing the English language proficiency needed to succeed in English-medium classrooms. Federal legislation refers to the term *limited English proficient (LEP)* to describe the same group of people.[2] According to Section 9101 of Title IX, an LEP student:

- is between the ages of 3 and 21;
- is enrolled or preparing to enroll in an elementary school or secondary school;
- has one of these three profiles:
    - Was not born in the United States or speaks a native language other than English
    - Is a Native American, an Alaska Native, or a native resident of the outlying areas, and comes from an environment where a language other than English has had a significant impact on his or her level of English language proficiency
    - Is migratory, has a native language other than English, and comes from an environment where a language other than English is dominant
- has difficulties in speaking, reading, writing, or understanding the English language that are so severe as to deny the individual one of the following:
    - The ability to meet the state's proficient level of achievement on state assessments described in section 1111(b)(3) of the NCLB Act
    - The ability to successfully achieve in classrooms where the language of instruction is English
    - The opportunity to participate fully in society

[2] Different terms have been used over the years for students whose second language is English. The term English language learner is in increased use since it more accurately represents the process of language acquisition.

Factors Influencing the Assessment of English Language Learners

This section describes factors to consider when developing assessments and making decisions regarding testing accommodations for ELLs. The factors are not guidelines per se, but rather provide useful context for the guidelines presented in the later parts of the document.

Language Factors

- *Different linguistic backgrounds*—ELLs in the United States possess a wide range of linguistic backgrounds. While the majority of ELLs come from Spanish-speaking backgrounds, it has been estimated that approximately 400 different native languages are spoken by ELLs nationally. This is particularly important to keep in mind when considering the use of native language testing accommodations, since it may not be possible to provide assessments in all native languages represented in a large school district or a state.
- *Varying levels of proficiency in English*—ELLs vary widely in their level of English language proficiency, and furthermore, ELLs may have varying levels of oral and written English proficiency. Do not assume that students who can converse easily in English will have the literacy skills necessary to understand the written directions for a standardized test. Some ELLs may be proficient in the English used for interpersonal communications but not in the academic English needed to fully access content-area assessments. Studies show that the level of language proficiency has an influence on processing speed. In other words, compared with native speakers, ELLs generally take longer on tasks presented in English. This is important to keep in mind when designing and scoring the assessment, as well as when making decisions about testing accommodations.
- *Varying levels of proficiency in native language*—ELLs also vary in their levels of proficiency and literacy in their native languages. Therefore, do not assume that speakers of other languages will be able to understand written test directions in their native languages. In fact, a large proportion of ELLs were born in the United States and may not have had any formal schooling in their native language. This is important to keep in mind when considering the use of native language accommodations.

Educational Background Factors

- *Varying degrees of formal schooling in native language*—As mentioned previously, ELLs vary widely in the level of formal schooling they have had in their native languages. The degree of native-language formal schooling affects not only native language

proficiency—specifically, literacy in the native language—but also the level of content-area skills and knowledge. For example, students from refugee populations may enter the U.S. educational system with little or no formal schooling in any language. These students must learn English and content-area knowledge simultaneously, while also being socialized into a school context that may be extremely unfamiliar. Other ELLs may come to the United States with more formal schooling and may have received instruction in the content areas in their native languages. The primary challenge for these students is simply to transfer their existing content knowledge into English. Again, these factors come into play when making decisions about appropriate accommodations.

- *Varying degrees of formal schooling in English*—ELLs also vary in the number of years they have spent in schools where English is the language of instruction. A distinction may also be made between students who have studied English as a foreign language while in their home countries and students who have studied English as a second language only in the United States. Furthermore, ELLs differ in the type of instruction they have received while in English-speaking schools. *Bilingual*, *full English immersion*, and *English as a second language* are but three of the many existing instructional programs for non-native English speakers, and there are great variations in how these programs are implemented. In addition, ELLs from migrant populations may spend many years in English-speaking schools but may also experience repeated interruptions and relocation to different cities in the United States in the course of their schooling, which may have an impact on both their English language proficiency and on their content-area knowledge.
- *Varying degrees of exposure to standardized testing*—It should not be assumed that all ELLs have had the same exposure to the standardized testing that is prevalent in the United States. Students in some countries may have had no exposure to multiple-choice questions, while those from other countries may never have seen a constructed-response question. Even ELLs from educationally advantaged backgrounds and with high levels of English language proficiency may not be accustomed to standardized, large-scale assessments and may be at a disadvantage in these testing situations.

Cultural Factors

Cultural factors can also be potential sources of construct-irrelevant variance that add to the complexity of appropriately assessing ELLs.

- *Varying degrees of acculturation to U.S. mainstream*—ELLs come from a wide range of cultural backgrounds, and cultural differences may place ELLs at a disadvantage in a standardized testing situation. Lack of familiarity with mainstream American culture, for example, can potentially have an impact on test scores for ELLs. Students who are unfamiliar with American culture may be at a disadvantage relative to their peers because they may hold different assumptions about the testing situation or the educational environment in general, have different background knowledge and experience, or possess different sets of cultural values and beliefs, and therefore respond to questions differently. Students from cultures where cooperation is valued over competition, for example, may be at a disadvantage in those testing situations in the United States where the goal is for each individual student to perform at his or her best on his or her own. Students from economically disadvantaged backgrounds may also respond to questions differently and may have background knowledge and experiences that are different from those presumed by a test developer.

Planning the Assessment

In planning assessments to be taken by the general student population, including ELLs, the general principles of good assessment practices apply. This section describes different steps within the planning process, highlighting issues most relevant to the assessment of ELLs.

Test Purpose

The purpose of a test must be clear in order for valid interpretations to be made on the basis of the test scores. Tests have different purposes. For example, one test may be used to evaluate students' readiness to advance to the next grade, while another evaluates students' need for remediation. It is also important to outline the specific interpretations that will be made based on the scores. For example, tests used as a criterion for high school graduation will affect students differently than tests designed to inform instructional decisions.[3]

[3] For additional information on the components that one should consider in test planning, please refer to the *ETS Standards for Quality and Fairness*.

Defining the Construct

A second criterion for validity is a precise and explicit definition of the construct the test is intended to measure. For K-12 assessments, state standards underlie the test specifications. Sometimes other state documents, such as curriculum frameworks, may clarify knowledge and skills stated in the standards. When defining a construct for an assessment to be given to ELLs, consider in particular how English language skills interact with the construct. For example, when defining the construct for a mathematics test, consider whether it is intended to be a test of mathematics, in which case the test should require no or absolutely minimal English proficiency, or a test of the ability to do mathematics within an English-language educational environment, in which case the ability to comprehend word problems in English may be part of the construct. Similarly, those who define the construct should pay attention to how much of the vocabulary of the discipline in English is to be viewed as part of the assessment.[4] Defining English proficiency as part of a target construct for an assessment in mathematics or science is neither right nor wrong. It is essential, however, that these definitions be explicit. Furthermore, even if English proficiency is part of the construct, take care to define what level of English proficiency should be expected of students. When defining the linguistic demands to be included in the construct, make an effort to include professionals with backgrounds in educating ELLs.

Developing the Assessment Specifications

Assessment specifications define the test content and explain how that content will be assessed. Assessment specifications also provide a link between a state's content standards and the items or tasks that appear in a particular test. ELLs will likely constitute a significant portion of the population of many K-12 tests; therefore, considering ELLs during the initial development of assessment specifications is utterly important. The following points relevant to ELLs should be addressed when writing K-12 assessment specifications.

Domain of Knowledge and Skills

States are likely to have documented content standards for the subject area to be assessed. States may also provide performance standards and other documents that define the domain and their expectations for student achievement. Test developers should review these documents carefully and note the degree to which each standard calls for the ability to read, write, speak, or listen in English.

[4] Some disciplines use *everyday language* to refer to certain disciplinary concepts (e.g., the terms *energy* and *transfer* in physics), while *specific language* terms are used for other concepts (e.g., the terms *mitosis* and *metamorphosis* in biology). Keep this in mind when evaluating the degree of English language proficiency needed for a given subject area.

Share the results of this review with the educational agency and clarify the level of English proficiency that each standard implies. Educational agencies may not be aware of ambiguities in their content standards regarding this issue. Content standards are often developed by committees of experts focused primarily on the subject area. Defining expectations about the use of English, use of ELLs' first languages, and use of visual representations is important both to ensure an efficient development process and to gain educational agencies' confidence in the validity of an assessment.

Many states define expectations for test questions in detail in item specifications, as distinct from assessment specifications. The item specifications contain detailed notes about acceptable vocabulary, content limits, and focus for each of the state standards assessed. Develop—and have the state approve—item specifications before the assessment program's first content or bias and sensitivity reviews. Update details in the specifications when items are reviewed, with state approval.

Number and Types of Items or Tasks

In general, all other things being equal, tests with more items will supply more reliable scores.[5] Reliability refers to the extent to which scores obtained on a specific form of an assessment can be generalized to scores obtained on other forms of the assessment, administered at other times, or possibly scored by some other rater(s). Thus, as is true for all students, it is desirable to provide ELLs with multiple opportunities to show what they know and can do.

Some have posited that ELLs should have not only multiple opportunities, but also multiple *ways* to show what they know, and that assessment specifications should include a variety of item and response types that may lead to assessments on which ELLs are more likely to be able to show their strengths. For example, items with visuals, performance tasks, or oral responses are sometimes suggested as ways to allow ELLs to better demonstrate proficiency. However, in the literature base, there is no consistent agreement as to whether these varied item types are in fact beneficial. In addition, more items and more sets of directions may tax the reading ability of ELLs, as well as the rest of the examinee population. Lastly, educational agencies will always have limitations regarding time and costs and must decide what is realistic for a given testing program.

Therefore, we suggest making an effort to present the best options for task types that allow ELLs to show what they know and can do within the practical limits of the assessment program. Item tryouts, discussed in a later section, may be a way of exploring the use of different item types with ELLs.

[5] The phrase *all things being equal* is a crucial one. Adding more items will increase reliability if the new items have the same characteristics as the existing items in the test, i.e., the new items measure the same construct and are affected at the same level by the construct-irrelevant factors such as unnecessary linguistic complexity and cultural biases. However, if the new items are more linguistically complex, or are affected by other sources of biases differentially, then addition of the new items may even decrease the reliability of the test.

Relative Weights of Tasks and Skills

The weight of a task or content category is generally decided by the importance of the assessed task relative to the other tasks on the test and the degree to which the tasks tap content described in the state's standards. For more information, refer to the documented decisions made during the process described under *Domain of Knowledge and Skills* to determine possible weightings. Often tasks that require more time to complete (and usually longer responses written in English) receive more weight in an assessment. Such weightings may disadvantage ELLs; therefore, develop a careful rationale for weighting to apply to all students' responses, taking both content knowledge and language skills into account.

Assessment and Response Forms

Assessment specifications describe how the tasks will be presented to the students and how the students are expected to respond. Printed test booklets and answer sheets on which students mark responses and write constructed responses are very common in the K-12 school environment. Just as including a variety of item types in an assessment provides multiple ways for ELLs to show their knowledge, some feel that incorporating different types of media (such as video or sound) in an assessment's presentation format may also benefit ELLs. However, the research base is not yet well developed on this topic, so use caution in employing different types of media. In addition, using alternative media may unintentionally disadvantage other groups of students, including students with disabilities such as visual impairments. Alternative forms of responding, such as using diagrams or tables, may help some ELLs—as well as students with different learning styles—better demonstrate what they know.

Just like students in the general population, ELLs vary greatly as individuals. Therefore, no one type of presentation or response is optimal for all ELLs. However, in general, keep in mind while developing assessment specifications that, depending on the content area being assessed, large amounts of text make it less likely that ELLs will understand what is being asked of them.

Some testing programs also rely on tasks that require extended written responses to assess students' depth of knowledge in the content areas. Where feasible, consider including tasks that allow examinees to respond in ways that do not require long responses written in English, such as by drawing a diagram or other visual representation, as appropriate. Also consider using item tryouts as a means of obtaining information on ELLs' responses to and performance on different kinds of tasks.

Cultural Background and Diversity

The educational agency for which an assessment is developed should be able to provide information about the cultural backgrounds of its test-taking population, including ELLs. Content standards may also refer to exposure or knowledge about cultural or regional history or literature. If possible, test material (e.g., item and stimulus material) should include references to and contributions of major groups in the tested population (see the *ETS Fairness Review Guidelines* for further information on representing diversity in test material). Discuss with the educational agency the ways in which cultural diversity is represented in passages, context setting, and illustrations. Test specifications should describe the type of material in each test form, and item specifications should describe the appropriate material for each standard.

Developing Test Items and Scoring Criteria

Matching the Task to the Purpose

The first step in developing a test item should be to link, directly to the test specifications and content standards, the content and skill that the item is supposed to measure. If the items require a high level of English proficiency, unrelated to the construct as defined, this will likely affect the scores for ELLs as well as students in the general population. For content area assessments, only include items that require high degrees of English proficiency if they are consistent with the assessment specifications. Examples of items that require a high degree of English proficiency are those that ask examinees to identify or provide specific definitions or terminology in English that are unrelated to the construct, or items that are evaluated based on the quality of the language in a constructed response.

Item writers and reviewers should work to ensure that all test items maintain specificity in their match to content guidelines. As part of the process of creating and reviewing test material to ensure that it is appropriate and accessible to examinees, it is important that item developers, state content review staff, and state review committees analyze each item critically to ensure that it only measures the intended construct.

Defining Expectations

Because ELLs—just like students in the general population—come from a wide variety of cultural and educational backgrounds, item writers should not assume that students have had any previous experience with given tasks. For example, students should be told explicitly what type of response is acceptable for a constructed-response question, whether it is a paragraph, complete sentence, list, diagram, mathematical equation, and so on. Likewise, the criteria for the evaluation of

the response should be made clear to the student. As this may add a significant reading load to the directions, information about how responses will be scored may be especially helpful if students receive it prior to the test.

Writing Appropriate Directions

Design directions to maximize clarity and minimize the potential for confusion. Consider options for simplifying the language used for directions (see below). Also consider presenting the directions orally or in a language other than English if that will provide the best, most understandable instructions for ELL examinees (see *Testing Accommodations*).

Using Accessible Language

Using clear and accessible language is a key component of minimizing construct-irrelevant variance. However, do not simplify language that is part of the construct being assessed (e.g., the passages on a reading comprehension test or challenging vocabulary that is part of the construct of a subject area test). In other cases, though, the language of presentation should be as simple and clear as possible. Some general guidelines for using accessible language are provided below:

- Use vocabulary that will be widely accessible to students. Avoid colloquial and idiomatic expressions, words with multiple meanings, and unduly challenging words that are not part of the construct.
- Keep sentence structures as simple as possible to express the intended meaning. For ELLs, a number of simple sentences are often more accessible than a single more complex sentence.
- Avoid use of negatives and constructions utilizing *not* in the questions' stems and options as they can cause confusion, especially for ELLs.
- When a fictional context is necessary (e.g., for a mathematics word problem), use a simple context that will be familiar to as wide a range of students as possible. A school-based context will often be more accessible to ELLs than a home-based context.

Ask reviewers to note any instances where an item can be simplified or clarified to make the language more accessible. However, do not change language that is part of the construct being measured.

Presentation

For all assessments, test developers should be aware of formatting issues. Fonts, font sizes, line breaks in paragraphs, and test directions should all receive a careful review. ELLs who already have reading ability in another language may have different levels of familiarity with texts that read from left-to-right, right-to-left, or top-to-bottom. Therefore, clearly and consistently placing elements such as pictures, page numbers, and other page elements can greatly improve readability for ELLs as well as other students.

Fairness and Sensitivity

In order to maximize fairness and accessibility for all students, the *ETS Standards for Quality and Fairness* (and, as applicable, the *ETS International Principles for Fairness Review of Assessments*) require that test materials "minimize the effects of construct-irrelevant knowledge or skills" and "avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting." In applying these guidelines, it is important to recognize that ELLs have had extremely diverse life experiences and may be unfamiliar with many U.S. cultural contexts. One way to increase accessibility for ELLs is to use school-based contexts for test items as often as is practical. For example, research has shown that mathematics word problems are more accessible for ELLs when set in a school context (e.g., counting things such as notebooks, desks, and erasers) than when set in a home context (e.g., counting the number of appliances in the home). Other neutral contexts and topics may be appropriate, as well; consider all available information about the test-taking population.

External Reviews of Test Materials

Reviews from diverse, informed points of view are an effective technique to improve the quality of assessments, including the degree to which assessments are accessible to ELLs. The insights external reviewers provide can help test developers understand how students are likely to interpret test materials and how members of different populations may respond to test items. Although it is expected that all test material will receive thorough internal reviews, external reviewers who are chosen for their knowledge of the ELL population and the specific challenges they face may be able to provide insights that complement and improve the work of the internal reviewers. This helps to ensure that the contexts selected for items and the language in which they are written are appropriate for ELLs.

The educational agency developing the test almost always requires that the materials pass its own committee reviews. An agency may also seek recommendations regarding the types of professionals it should invite to review an assessment. If a test will be administered to ELLs, the review panels should include, in addition to content experts, professionals who are familiar with issues regarding

different ELL populations, such as migrant, newly arrived, or reclassified students. In their reviews, the panels should also consider the variety of programs of English language instruction experienced by the students. The panels should, within the context of the state standards and the item specifications, evaluate each item for technical quality, alignment to standard, and accessibility to ELLs. The panels should also include in such reviews test specifications, directions, sample items, and scoring criteria.

External reviewers should address the following questions:

- Does each task match the purpose of the assessment and the assessment specifications?
- Are the directions for each task clear and appropriate?
- Is the task presented in clear and accessible language, free from idioms and complex linguistic constructions?
- Are the formats of both the assessment and the response materials appropriate?
- Do the tasks and scoring criteria meet standards for fairness?

Evaluating the Tasks Through Tryouts

Trying out or field-testing items can provide extremely useful information during the test development process. When conducting an item tryout, use a sample of examinees similar to those who will take an assessment once it is administered operationally (for official score-reporting purposes). This step is particularly important for items that will be used with ELLs.

Purposes of Item Tryouts

There may be several reasons to conduct item tryouts. Data may be collected in order to:

- inform decisions about how appropriate the items are for a sample of examinees similar to the operational population,
- inform content and fairness reviews of the items,
- evaluate timing requirements for new or existing item types,
- evaluate the clarity of instructions to examinees,
- support the scaling or equating of test forms,
- inform the standard setting process by providing performance data, which panelists will receive as feedback on cutscores, on different groups, and
- assess whether ELLs of different proficiency levels can understand the text of the items. This is important when English language proficiency is not the construct of interest.

Types of Item Tryouts

Item tryouts may take several different forms, ranging from one-on-one interviews with students, through small-scale pilot tests, to large-scale field tests. As with other activities described within these guidelines, it may not be possible to implement each of these types of item tryouts in a given testing program because of resource constraints. However, we describe them here so that readers can make informed decisions about when and whether each type may be useful.

One-on-One Interviews

One-on-one interviews with students who have been administered the items can provide much useful information. These interviews can take the form of informal debriefings after students have completed the tasks, or more formal cognitive laboratory activities where students are interviewed either while they are answering the questions or afterward.

Because individual interviews are time-consuming to conduct, it is usually not possible to involve large numbers of students. The information that such interviews yield can sometimes be idiosyncratic. However, the quality and type of information interviews provide can offset that concern. Interviews allow students to talk about the cognitive processes they employed when answering the item, whether anything confused them, and how they arrived at their answer. The interviewer can also ask students what they think the item is asking them to do or what they think the item is measuring. Qualitative summaries of this feedback can be very helpful for the item review process. This type of item tryout is particularly important for items that will be used with ELLs, since the interviewer can ask them directly about their understanding of and response to the items.

For ELLs, interviews are extremely useful for identifying potential threats to the validity of tests that measure knowledge in content areas other than English language arts. To determine whether items require a high degree of English proficiency unrelated to the construct, it is important to assess ELLs' understanding of the language of the items. While external reviewers with expertise in ELL issues can provide valuable insights, working directly with ELLs to gather their impressions of test materials can generate even more detailed and useful information.

Since one-on-one interviews may be costly and time-consuming, it may not be possible to conduct them as part of an ongoing testing program. They may be most useful when trying out a new item type. Testing officials will need to decide whether the information they may gain from these interviews is worth the time and expense.

Small-Scale Pilot Tests

Small-scale pilot tests may also provide useful information on how students respond to the items. In this data collection format, test developers administer the items to a larger sample of students than is used for one-on-one interviews, and, generally, one-on-one debriefing does not take place. Because these samples may not be fully representative of the test-taking population, the item statistics provide only a gross measure of whether students were able to answer the item correctly. Including a small-scale pilot with an oversampling of ELLs may prove very helpful during the item development process to discover issues specific to ELLs. Again, however, budgets and schedules may not allow for these types of pilot tests to take place. Such activities may be most appropriate when introducing a new item type.

Large-Scale Field Tests

In large-scale field tests, test developers administer the items to a large, representative sample of students. Because of the size and nature of the sample, statistics based on these responses are generally accurate indicators of how students may perform on the items in an operational administration. If the tryout items are administered separately from the scored items, motivation may affect the accuracy of the results. When the tryout items are embedded among the scored items, students do not know which items count and which do not, so motivation is not a factor. Consequently, many states conduct embedded field testing and are increasingly moving toward placing the tryout items in random positions within each test form. Conducting a large-scale field test on a group in which ELLs are well-represented will allow for the evaluation of item difficulty and other item characteristics specific to ELLs.

Guidelines for Item Evaluation

The type of tryout should be tied to the goals of the evaluation. To obtain information directly from students about their thought processes while answering the items, conduct one-on-one interviews. To obtain information directly from ELLs about their understanding of complex language in items measuring content areas other than English language arts, conduct one-on-one interviews. Evaluate the extent to which complex language generates comprehension difficulties for ELLs relevant to the construct being measured. If there appears to be unnecessary linguistic complexity, review the item and revise it as appropriate before the operational administration. Field test it again if necessary (for example, in the case of pre-equated tests).

To inform judgments about how items will work, conduct a small-scale pilot test—but remember that the data from such pilots usually does not come from a representative sample. To

obtain reliable and valid statistics that can be used when selecting items for test forms or equating, conduct a large-scale field test.

Try items out on a sample that is as similar as possible to the population that will take the operational administration. However, oversampling ELLs during pilot testing is recommended; such oversampling increases the likelihood of uncovering issues that may be specific to those students. Document the procedures used to select the sample(s) of examinees for item tryouts and the resulting characteristics of the sample(s).

Try out all item types, including both selected-response, constructed-response, and hands-on tasks or activities. If constructed-response items are tried out, score them using scorers and procedures that are as similar as possible to those used for operational administrations (but consider possible security risks engendered by exposing prompts before the administration). Evaluate responses to constructed-response items according to the following criteria, per the *ETS Guidelines for Constructed-Response and Other Performance Assessments*:

- Do the examinees understand what they are supposed to do?
- Are the tasks appropriate for this group of examinees?
- Do the tasks elicit the desired kinds of responses?
- Can the responses be easily and reliably scored?
- Can they be scored with the intended criteria and rating scale?
- Are the scorers using the scoring system in the way it was intended to be used?

To ensure accessibility for ELLs, it is also important to ensure that rubrics focus on the construct of interest and do not include construct-irrelevant variance by placing inappropriate emphasis on English language proficiency unrelated to the construct. For example, scoring rubrics should state clearly that, when English language proficiency is not defined as part of the construct, raters should ignore errors in English when scoring for content. For more information, see *Scoring Constructed-Response Items*.

Limitations of Item Tryouts

Even when field test samples and operational populations seem comparable, differences in demographics, curriculum, and culture may make comparisons difficult. Document any limitations of the representativeness of the field test sample. Such limitations are most likely to be present for one-on-one interviews and small pilot samples. Motivational level can also be a factor as field test participants are often not as highly motivated to do their best as are operational examinees. In sum,

field testing is valuable for trying out new tasks and scoring criteria, but use the results of field testing with caution for higher-stakes decisions such as setting the standard for passing the assessment.

Scoring Constructed-Response Items

While issues related to scoring apply to the general population, scoring constructed responses written by ELLs may present a number of additional unique challenges. At first glance, constructed responses from ELLs may be confusing to read and may appear to be off-topic or unscorable. Many of these responses, however, can be scored—and can possibly receive high scores—if the scorer has been trained to identify and properly evaluate the multiple ways an examinee might approach an item. Two important ways in which ELLs' constructed responses may differ from those of other students are differences due to language background and differences in the style of the response.

Differences due to language background will vary among students, but some patterns are generally recognizable. ELLs may use spelling conventions or false cognates based on their knowledge of their first language. They may spell phonetically or mix words or word parts between English and another language. Other frequent markers of ELL responses may be missing articles, lack of noun/verb agreement, or incorrect use of prepositions. Scorers may also find that ELLs combine words that should not be combined. ELL responses may also be characterized by sentence patterns that reflect reasoning patterns used in the test taker's native language. While it is appropriate to consider these types of errors in a test of English-language writing skills, raters should overlook them in tests of academic content knowledge. Again, define the construct as explicitly as possible so that raters can differentiate construct-relevant factors from construct-irrelevant ones.

ELLs may also differ considerably from native English speakers in the style in which they present constructed responses. For example, ELLs that have learned long division in another country may show their work moving from the bottom of the page to the top or using other conventions for long division, unlike the common practice in the United States of writing out long division and remainders moving from the top of the page to the bottom. ELLs may also attempt to communicate their answers in alternate ways, such as by drawing diagrams and pictures.

Of course, whether any of these responses or styles is acceptable depends on the test and the construct being measured. While an ELL's lack of control of fundamentals such as sentence structure or word order may appear to indicate that he or she has responded to an item poorly or incorrectly, scorers should recognize that some aspects of an ELL's response may only show unfamiliarity with English and not low proficiency in the construct. Including scorers and scoring leadership (such as table leaders) who are familiar with the teaching and learning of ELLs in the process of scoring can help scorers who come across unfamiliar or confusing responses from ELLs.

We are not suggesting that responses that appear to have been written by ELLs be routed to scorers familiar with ELL issues, since that may introduce bias into the scoring process. Similarly, we are not necessarily recommending that response issues common to ELLs be identified as such, since that could also potentially bias scorers. Instead, we recommend describing these issues in more general terms to all scorers as reflective of all students who lack mastery in English language writing conventions.

The *ETS Guidelines for Constructed-Response and Other Performance Assessments* outline general steps that should be taken in the scoring process: creation of rubrics, recruiting scorers, training scorers, and confirming consistent and accurate scoring. Each of these steps has specific application to scorers who will evaluate ELLs' responses, as discussed below.

Creation of Rubrics

For content area assessments, the scoring leadership should examine constructed-response items and determine whether they require specific English-language terms or constructions in order to receive a high score. For example, if the test specifications require examinees to be able to define key terms in English and use them in a response, then a certain level of English proficiency is, in fact, part of the construct. If, however, the test specifications require that the student be able to describe or represent things such as a scientific process or mathematical function, then specific terms and usage in English may not be required to receive a high score.

After determining the extent to which specific English language skills are required for answering an item, write rubrics so that raters can interpret responses in a linguistically sensitive way. That is, the rubrics should make clear the role that English language skills should play in determining a score. (It may be helpful to have educators who are familiar with the performance of ELLs involved in the creation and review of rubrics). Generally, write rubrics for content area tests so as to focus on content rather than on language use—but carefully evaluate the construct to determine if, for example, writing an essay in English to provide evidence about a historical event would in fact require a certain degree of language skills. For assessments of English writing skills, the rubric should consider command of language (vocabulary, grammar, mechanics, etc.) but also make clear the role of critical thinking as distinct from fluency in English-language writing conventions. While this is not an easy distinction to make, it is an important consideration. Rubrics should be clear about how raters should score responses written partially or entirely in a language other than English. That determination should also be made clear to students in information distributed about the test beforehand.

Recruiting Scorers

The proper scoring of ELLs' responses includes an understanding of the language or presentation style examinees use. Knowledge of second language acquisition, ELL teaching background, or other aspects of cultural background may help raters to appropriately evaluate some responses ELLs produce. Including in the group of scorers (and scoring leadership, such as table leaders) people who are familiar with aspects of responses that have characteristics of students learning English as a second language can help to ensure more accurate scoring for ELLs. These scorers could serve as resources when ELL-related issues arise. To reiterate, we are not suggesting that responses that appear to have been written by ELLs be routed to those scorers, since that may introduce bias into the scoring process.

Training Scorers

Scorer training should include a review of how to interpret responses and the scoring rubric in a linguistically sensitive way. Training should make clear the extent to which particular responses must contain key terms or other specific language in English in order to be considered for the top scores. Assessment developers and chief readers/table leaders should pick out exemplar responses, at various score points, that evince some or all of the ELL characteristics noted above, including some that are presented in atypical formats. These exemplars, in tandem with the rubrics, should be used in training raters. Through these exemplars (and the explanations that go along with them) raters can be trained to recognize ELL characteristics and to score ELL responses fairly without introducing bias. Scorers-in-training should receive an explanation of the extent to which the examinee's level of English proficiency affected the scoring. Low levels of English proficiency can affect the scores of many students, not just ELLs. As with all scoring, instructions should tell scorers how to handle responses written entirely in languages other than English.

Confirming Consistent and Accurate Scoring

Using training papers that reflect characteristics of ELLs' responses can help scorers become familiar with the rubric and how it applies to a range of responses. All aspects of scorer training—both before scoring begins and while it is ongoing—should include responses by ELLs (if they can be identified) as part of the training materials. Recalibrating scorers at the beginning of each scoring session should confirm scorers' abilities to resume accurate scoring. Including ELLs' responses as calibration papers (given at the start of a scoring session) and as monitor papers (embedded among other student responses while scoring is underway) is an effective means of confirming scorers' use and interpretation of a rubric at any point in time. The scoring leadership should confirm the validity

of all sample student responses used in training. It is beneficial to include among the scoring leaders professionals who are knowledgeable about English language learning.

Testing Accommodations for English Language Learners

Purpose of Testing Accommodations for English Language Learners

The main purpose of providing examinees with testing accommodations is to promote equity and validity in assessment. For ELLs, the primary goal of testing accommodations is to ensure that they have the same opportunity as students who have English as their first language to demonstrate their knowledge or skills in a content area. Reducing or eliminating construct-irrelevant variance from the testing situation increases the likelihood that score users will be able to make the same valid interpretations of ELLs' scores as they make for other examinees. In general, the main sources of construct-irrelevant variance on content area assessments for ELLs are the effects of English language proficiency in answering test items. Unless language proficiency is part of the construct being measured, it should not play a major role in whether an examinee can answer a test item correctly.

Accommodations refer to changes to testing procedures, which researchers have traditionally considered to include presentation of test materials, students' responses to test items, scheduling, and test setting. As a general principle, testing accommodations are intended to benefit examinees that require them while having little to no impact on the performance of students who do not need them. At present, the research basis regarding which accommodations are effective for ELLs under what conditions is quite limited. Relative to research on students with disabilities, research on accommodations for ELLs has a much shorter history, with the results from studies often seeming to contradict each other.

Some state policies distinguish between *testing accommodations* (changes in the assessment environment or process that do not fundamentally alter what the assessment measures) and *testing modifications* (changes in the assessment environment or process that may fundamentally alter what the assessment measures) and refer to both as *testing variations*. In these guidelines, the term *testing accommodation* refers to changes that do not fundamentally alter the construct being assessed.

Identifying Students Eligible for Accommodations

Policies for identifying ELLs who may be eligible for testing accommodations continue to evolve. At present, there are no uniform guidelines or policies at the federal level regarding the use of accommodations for ELLs. For students with disabilities, eligibility for accommodations is part of a student's Individualized Education Plan (IEP); however, ELLs do not have any corresponding documentation. Across states and local school districts, both the eligibility requirements as well as the

specific accommodations available to ELLs vary widely. In fact, some policies are not transparent with respect to how eligibility for accommodations is determined or who is making the decisions for ELLs. As a general principle, if an ELL's English language proficiency is below a level where an assessment administered in English would be considered a valid measure of his or her content knowledge, then that student may be eligible for one or more testing accommodations.

Typically, ELLs who regularly use accommodations in the classroom are usually eligible to use the same accommodations in testing situations. However, some accommodations that may be appropriate for instruction are not appropriate for assessment. For example, some ELLs routinely have text read aloud to them as part of instruction. But if decoding or reading fluency is being assessed as part of reading comprehension, this would not be an appropriate accommodation because it would change the nature of the assessment from one of reading comprehension to one of listening comprehension. Further, an accommodation such as the use of a native language glossary of terms that could be appropriate for certain subjects such as mathematics or science would not be appropriate for English language arts, because the use of a glossary would change what is being assessed and would provide an unfair advantage to those who have access to it.

Identifying Accommodations

Testing accommodations for ELLs can be broadly grouped into two categories: Direct linguistic support accommodations (which involve adjustments to the language of the test) and indirect linguistic support accommodations (which involve adjustments to the conditions under which a test is administered). To be ELL-responsive, an accommodation should provide some type of linguistic support in accessing the content being tested.

To date, the limited number of research studies on accommodations for ELLs indicates that direct accommodations appear to benefit student performance more than indirect accommodations. Examples of direct linguistic support accommodations include providing a translated or adapted version of the test in the student's native language or providing test directions orally in the student's native language. The use of translated tests is a complex issue because questions can arise as to whether the original and translated versions are measuring the same construct in the same manner. Translated versions of items may or may not have the same meaning as in their original versions. Therefore, some educational agencies have created *transadapted* versions of tests, which are translated versions of tests that have been culturally adapted for the examinees. Furthermore, the use of translated tests may only be of limited benefit to examinees, particularly if the language of instruction and the language of the test are not the same. Furthermore, unless a test can be translated into all of the native languages spoken by the students in a school district or state, questions of equity may arise

if translations are available only for a limited number of languages. In addition, in some states, public policy may prohibit the assessment of students in languages other than English.

Examples of indirect linguistic support accommodations include extended testing time or having the test administered individually or in small groups. Some of these accommodations do not address construct-irrelevant variance due to language; however, they may be useful or necessary to facilitate test administration for ELLs or for all students. Because state and local policies are evolving at a rapid pace, we have not provided with these guidelines a complete list of accommodations that state or local school districts allow for ELLs. Test developers and interested readers should contact the appropriate educational agencies to obtain the most current assessment policy and list of accommodations available to ELLs.

Some states have simply extended to ELLs the use of accommodations originally intended for students with disabilities. However, some of these accommodations are clearly inappropriate when applied to ELLs (such as the use of large print versions of tests, which are appropriate only for students with a relevant disability such as a visual impairment). Recent reviews indicate that fewer than two thirds of the accommodations for ELLs found in states' assessment policies address the unique linguistic needs of ELLs exclusively.

When Accommodations Should Be Used

At present, there are no existing standards that can definitively guide the use of testing accommodations for ELLs. The appropriate use of accommodations depends on a number of factors including: a student's proficiency in English as well as his or her native language, the academic subjects being assessed, the student's familiarity with the accommodations, the language in which the student receives instruction, and the range of available accommodations for examinees. To the extent practical, decide on accommodations for individual students, not as a collective group. The accommodation or combination of accommodations that may be most appropriate for one ELL may or may not be the best choice for another student.[6] Within the past decade, some progress has been made in developing systems for making decisions on testing accommodations for ELLs, but additional work is necessary before any of these systems are ready for use by administrators or teachers.

Currently, without sufficient research findings to inform appropriate use of accommodations for ELLs, accommodation decisions are best guided by the following operating principles: Most importantly, accommodations for ELLs should not alter the construct being assessed; this is particularly critical when students are tested on their academic content knowledge and skills. In

[6] Status as an ELL is much more dynamic than disability status or cognitive status, and a student's ELL proficiency level may change from one year to the next. For this reason the student's need for a given accommodation may change from one year to the next due to increased English language proficiency.

addition, the choice of accommodations should allow ELL examinees to demonstrate their knowledge and skills to the greatest extent possible. This means ELLs should receive the greatest degree of linguistic support accommodations—such as a glossary or bilingual dictionary—necessary in order to ensure this outcome.

Using Statistics to Evaluate the Assessment and Scoring

Multiple sources of empirical evidence should be gathered to evaluate the fairness of assessments and scoring.[7] The *ETS Standards for Quality and Fairness* state that, whenever possible and appropriate (i.e., if sample sizes are sufficient), testing programs should report analyses for different racial/ethnic groups and by gender, and that testing programs should use experience or research to identify any other population groups to be included in such evaluations for fairness. Therefore, we recommend that in K-12 assessments, testing programs should, where possible, report disaggregated statistics for native English speakers, ELLs, and former ELLs, so that the distributions of scores for these groups can be evaluated. Programs should also review differences in scores across testing variations (types of accommodations and test modifications). Whenever appropriate, programs should report analyses for test variations commonly employed with ELLs. These include:

- language of assessment, translated versions of the test or dual language booklets (e.g., English vs. Spanish),
- linguistically modified (or plain English) versions of tests, and
- extended time, reading aloud instructions, and use of bilingual glossary.

Differential Impact

For each studied group (or test variation, if appropriate), the following statistical information can provide evidence regarding the validity of an assessment for different examinee groups:

- *Performance of studied groups.* Provide statistics about the performance of studied groups on the whole test, subtests, and items. Group differences in the distribution of scores and item and test statistics are worthy of investigation in order to determine the underlying causes of these differences.
    - For the test and, if appropriate, for subtests, compute score distributions and summary statistics—means, standard deviations, selected percentiles (the 10th, 25th, 50th, 75th, and 90th)—and percentages of students in each achievement level.

[7] This section assumes familiarity with psychometric and statistical concepts.

  - For individual items, report item difficulty, item-test correlations, and item characteristic curves.

- *Differential item functioning (DIF).* Report DIF statistics, if sample size allows, using ELLs as the focal group and non-ELLs as the reference group. If sample sizes allow, DIF results could also be reported using former ELLs as the focal group. Examine test items that are flagged as exhibiting DIF against one or more examinee groups in order to identify the possible causes, which can be useful in making decisions about possibly removing items from scoring.
- *Differential predictive validity.* Report statistical relationships among reported scores on tests and subtests and criterion variables (such as scores on other tests given in later years) for ELLs and non-ELLs. Gather information about differences in prediction as reflected in regression equations, or differences in validity evidence for studied groups. Evidence of differential predictive validity indicates that the test functioned differently for different examinee groups and suggests that further investigations into the construct validity of the test for all groups may be warranted.

Reliability

To investigate whether scores are sufficiently reliable to support their intended interpretations, the following statistics for each of the examinee groups are particularly informative:

- If sample size permits, provide the following for reported scores, subscores, and cutscores (if available): Reliability estimates (accounting for a variety of sources of measurement error), information functions, index of classification consistency (consistency of the pass/fail decisions based on cutscores), standard error of measurement (for raw and scaled scores), and conditional standard errors of measurement around cutscores.
- When comparing test reliability across studied groups, evaluate differences in group dispersion (for example, ELLs may be more homogeneous than non-ELLs). If reliability coefficients are adjusted for restriction of range, provide both adjusted and unadjusted coefficients.
- For scoring constructed responses, follow the *ETS Guidelines for Constructed Response and Other Performance Assessments* (i.e., estimate inter-rater reliability for individual items). Since ELLs' writing skills in English are in most cases lower than those of English-proficient

students, evaluate whether there are interactions between rater scoring and ELL membership.

Validity

The *ETS Standards for Quality and Fairness* recommend gathering evidence about whether a test is measuring the same construct(s) across different subpopulations. These standards also indicate that, if the use of an assessment leads to unintended consequences for a studied group, the testing program should review validity evidence to determine whether the consequences arose from invalid sources of variance—and, if they did, revise the assessment to reduce, to the extent possible, the inappropriate sources of variance.

For ELLs as well as non-ELLs, some methods for investigating validity include:

- *Analyses of internal test structure.* Report statistical relationships among parts of the assessment (e.g., intercorrelations among subtests, item test correlations, dimensionality and factor structure).
- *Relations to other variables/constructs.* Report statistical relationships among reported scores on the total test and subtests and with external variables.
- *Test speededness.* Because of ELLs' lower reading fluency, test time limits may affect their performance disproportionately relative to non-ELLs. For timed tests, evaluate the extent to which there are differential effects of test speededness on ELLs. Report the number of items not reached and omitted for each examinee group.

Summary

The purpose of these guidelines is to provide practitioners with a framework to assist in making appropriate decisions regarding the assessment of ELLs in academic content areas. These guidelines offer recommendations on many important assessment issues regarding ELLs, including the development of assessment specifications and items, reviewing and field testing items, scoring of constructed responses, test administration, testing accommodations, and the use of statistics to evaluate the assessment and scoring. Although the research literature is limited and does not yet provide answers to many issues related to the assessment of ELLs, we have based our recommendations on the most accurate information currently available, and we hope that test developers and other educators will find these guidelines to be helpful in improving the assessment and education of all ELLs. We also recommend that research into the validity of assessments for ELLs continue in order to provide even sounder bases for recommendations in this area.

Bibliography

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8,* 231-257.

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Erlbaum.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4)*,* 36-46.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74,* 1-28.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219-234.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test.* New Haven, CT: Yale University Press.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Educational Testing Service. (2003). *ETS fairness review guidelines.* Princeton, NJ: Author.

Educational Testing Service. (2006). *ETS guidelines for constructed-response and other performance assessments.* Princeton, NJ: Author.

Educational Testing Service. (2007). *ETS international principles for fairness review of assessments.* Princeton, NJ: Author.

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English language learners in U. S. schools: Report and workshop summary.* Washington, DC: National Academy Press.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Kopriva, R. J. (2008). *Improving testing for English language learners.* New York: Routledge.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11-20.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, *78*, 333-368.

Martiniello, M. (in press). *Linguistic complexity of math word problems, schematic representations, and differential item functioning for English language learners* (ETS Research Report). Princeton, NJ: Educational Testing Service.

Rabinowitz, S. N., & Sato, E. (2006). *The technical adequacy of assessments for alternate student populations: Guidelines for consumers and developers*. San Francisco: WestEd.

Rivera, C., & Collum, E. (Eds.). (2008). *State assessment policy and practice for English language learners: A national perspective.* Mahwah, NJ: Erlbaum.

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653-673). Mahwah, NJ: Erlbaum.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*, 170-192.

Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies* (College Board Research Report No. 2008-6; ETS Research Report No. RR-08-48). New York: College Entrance Examination Board.

Notes

Notes

Notes

Listening. Learning. Leading.®

www.ets.org